The Application of Professionally Accepted Standards for Reliability and Validity to the Collection of Evidence

Prepared by C. Taylor and J. Willhoft

August 14, 2006

One of the three options that have been legislated as alternatives to performance on the Washington Assessment of Student Learning (WASL) as a means for students to earn a Certificate of Academic Achievement (CAA) is a collection of work samples, also referred to as the Collection of Evidence[1] (COE). Legislation requires that the guidelines and protocols for submission and the criteria used for scoring "meet professionally accepted standards for a valid and reliable measure of grade level expectations and the essential academic learning requirements." (SB 6475, Laws of 2006)

The process recommended by OSPI to the State Board of Education (SBE) is that the standards shown in Tables 1A and 1B, from the ***Standards for Reliability and Validity of Classroom-Based Assessments,*** be reviewed and approved by the National Technical Advisory Committee (NTAC). NTAC approval will assure the SBE that the criteria for reliability and validity against which the COE will be judged meet "professionally accepted standards". The review and approval of these reliability and validity standards will take place in two stages. First, the CAA Options Advisory Committee, composed of national and local educators and assessment experts (See Appendix A) will review, refine (as needed), and approve the standards. These standards will then be submitted to the NTAC for their approval in August of 2006. Once the NTAC adopts a set of reliability and validity standards for the COE, the design features of the COE will be submitted for their review. The NTAC will be asked to reach consensus on the

[1] Collections of Evidence are subject specific (i.e., reading, mathematics, and writing) collections of classroom-based assessments or work samples for individual students that demonstrate comparable curriculum standards as those assessed by WASL.

alignment of design features of the COE that address the standards. That work will be completed in mid-August, and will be presented to the SBE at its August meeting.

Table 1A: Validity Standards for Classroom-based Assessments

Validity Standard 1: **Representation and Fidelity**	Do the knowledge and skills required by the assessments represent the breadth of knowledge and skills defined in the standards?
Validity Standard 2: **Cognitive Demands**	Do the assessment tools and processes require students to demonstrate the targeted knowledge and skills at a cognitive level specified in the standards?
Validity Standard 3: **Consistency Across Assessments**	Do different assessments of the same knowledge and skills elicit comparable work?
Validity Standard 4: **Alignment with Instruction**	Does assessment align with the content taught and the instructional methods used?
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	Do the assessment tools and processes provide an equal opportunity for individuals, regardless of group or setting, to demonstrate the targeted knowledge and skills?
Validity Standard 6: **Consequences of the Interpretation and Use of Assessment Results**	Are there negative consequences for students that could be prevented if assessment tools, processes, events, or decisions had been more valid?

Table 1B: Reliability Standards for Classroom-based Assessments

Reliability Standard 1: **Generalizability**	Is the work typical of what the student knows and is able to do in relation to the learning targets?
Reliability Standard 2: **Sufficiency of Evidence**	Is there sufficient evidence so that one can make a dependable judgment about what each student knows and is able to do in relation to the learning targets?
Reliability Standard 3: **Clarity of Directions and Expectations**	Do the assessment directions provide clear, unambiguous expectations so that students can dependably demonstrate what they know and are able to do in relation to the learning targets?
Reliability Standard 4: **Quality of Scoring**	Are the scoring rules and scoring processes systematic enough to ensure consistent evaluation over time and across diverse samples of student work that demonstrate the same learning targets?

Two sources served as source materials for the attached *Standards for Reliability and Validity of Classroom-Based Assessments*. The first source was the *Standards for Educational and Psychological Testing* developed jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). The fourth edition of these standards was published in 1999. This document is widely accepted within the community of measurement professionals as encompassing the standards to be met for the development, evaluation, and use of tests that are commercially-developed or are used in large scale public assessment systems. The second source was Taylor and Nolen (1996, 2005), in which the authors adapted the *Standards* for application to the classroom assessment context. This latter work was used as the basis for the standards presented in the *Standards for Reliability and Validity of Classroom-Based Assessments*.

Considerations in Applying these Standards to Collections of Evidence

The Collections of Evidence (COE) to be used for the CAA involve the use of classroom-based assessments in a large-scale assessment context. The COE process requires students to collect work samples from classroom assignments and organize this evidence for a large scale purpose. In this case, not all standards for the validity and reliability of classroom-based assessments can be fully addressed by design features of a large scale assessment program. Three validity standards and one reliability standard for classroom-based assessments have limited applicability in this large scale context.

Validity Standard 4 (Alignment with Instruction) can best be evaluated by the classroom teacher or the students who know whether instruction has prepared the students to demonstrate the knowledge and/or skills required by the assessments.

Validity Standard 6 (Consequences of the Interpretation and Use of Assessment Results) requires ongoing research related to validity standards 1-5 and the consquences of the COE for students. Consequences related to students' self-concepts, their conceptions of school and the subject disciplines, and their academic choices as a results of their classroom-based assessment experiences are beyond the scope of the COE. However, consequences related to the COE should be examined. Positive or negative consequences that arise from decisions made based on the collections are relevant to validity ONLY if these consequences are due to problems related to validity standards 1 through 5.

In addition, although it is possible to Enhance Fairness and Minimize Bias (Validity Standard 5) through careful selection of collections to use for scorer training, it is difficult to thoroughly assess Validity Standard 5 without more information about the students. As with Validity Standard 4 (Alignment with Instruction), only the classroom teacher and the students know

whether the features of the assessment tools or events allow students to demonstrate what they know andare able to do. It is possible, however, to ensure that the COE provides opportunities for all qualified students, to demonstrate their knowledge and skills. The guidelines for the COE can be evaluated for the degree to which they enhance fairness and minimize bias.

Finally, for Reliability Standard 2 (Clarity of Expectations) the protocols for the COE, and any subsequent training materials and directions for teachers and students can be evaluated for clarity of expectations. The clarity of directions for assignments can be evaluated only if directions for assignments are provided along with students' work samples. Finally, if students include tests as part of their collections, test questions can be evaluated for clarity.

Above and beyond issues of reliability and validity, a separate standard has been recommended by the CAA Options Advisory Committee to answer the question: "Are there unintended consequences, for students, schools, and districts, of using the assessment system to make decisions about students?" This standard is important to consider when collections of evidence are used to judge students' proficiency in relation to the standards. Examples of unintended consequencies might include poor WASL performance due to the COE option (which would have implications for a school, district, or state AYP), a narrowing of the curriculum to a limited number of assessment tasks, repeated practice with a single task until the student prepares a proficient performance, or other unintended consequences. Studies should be planned to determine whether there are unintended negative consequences of the COE.

In Tables 2A through 2G of this document, the design features of the COE are more fully detailed. Tables 3A and 3B of this document present the approved links between the design features of the COE and the professionally accepted standards for reliability and validity from *Standards for Reliability and Validity of Classroom-Based Assessments*.

Table 2A:
Protocols – Directions to the COE users to indicate the types of evidence needed for each subject area

Writing Protocol
There are to be 5 to 8 written samples that together demonstrate proficiency in idea/development, organization, style, and the use of conventions. More work samples do not equate to a better score: Carefully selected work samples is a better indicator. Work samples should be written in blue or black ink or word processed. ➢ At least one expository or persuasive on-demand essay, timed and supervised in class ➢ At least two expository non-timed essays ➢ At least two persuasive non-timed essays ➢ 3 work samples (including the on-demand sample) may not include any adult assistance beyond setting the prompt and public expectations for an effective paper. ➢ Other work samples may include drafts read with teacher input and general comments (e.g., "You need to check for spelling errors." or "You need to rework your conclusion to wrap up your writing and give your reader something to think about.").
Reading Protocol
Work samples that cover all six strands that are assessed on the Reading WASL. ➢ A minimum of 8 and a maximum of 12 work samples from a classroom setting or a teacher-approved independent setting. Half of the work samples must represent responses to literary text and half of the samples must represent responses to informational text. ➢ All texts used in the work samples must meet high school expectations for rigor of reading material. The work samples must be comparable in rigor in skill and content to the High School Reading WASL. ➢ Work samples may feature work completed in other content areas—science, social studies, CTE coursework, etc. However, they must still address the literary or the informational strands listed above. ➢ One work sample must be a literary analysis paper of a significant piece of text—short story, narrative essay, novel, etc. that includes a demonstration of more than one literary strand. ➢ One work sample must be a research paper that includes at least two texts used for research purposes. Examples of this type of reading responses include: magazine or newspaper article analysis, analysis of historical events or scientific procedures, etc. The work sample should demonstrate more than one informational strand. ➢ One work sample that must be completed in an "on-demand" setting where students are provided an assignment to complete within a class period and without any teacher or peer assistance.
Mathematics Protocol
There must be 8 to 12 work samples. ➢ A variety of work samples such as projects, assignments, or exams ➢ Work samples of moderate or high complexity to ensure moderate or high level cognitive demands of the student ➢ At least two high school level work samples that and can be scored for an entire target from a strand of EALR 1: ➢ At least two high school level work samples can be scored for an entire target* from a strand of EALRs 2 through 5: ➢ Work samples that combine a content strand from EALR 1 and a process strand from EALRs 2 through 5. Work samples for EALRs 2 through 5 must be distributed across EALR 1 content strands. ➢ Work samples you select for EALR 1 should be representative of multiple High School WASL Mathematics Test Specifications ➢ Work samples you select must combine at least one content strand from EALR 1 and at least one process strand from EALRs 2–5. Work samples should be complex enough to demonstrate moderate to high level thinking skills.

Table 2B:

Sufficiency Review – Process used to determine that all of the WASL learning targets for a domain are included in the collection

Writing Protocol
In order to meet the sufficiency guidelines for successfully submitting a Writing Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Reading Protocol
In order to meet the sufficiency guidelines for successfully submitting a Reading Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Mathematics Protocol
In order to meet the sufficiency guidelines for successfully submitting a Mathematics Collection of Evidence, the student and teacher preparing the collection must comply with the following guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored

Table 2C:

Work Sample Documentation

Writing Protocol	*In the "Work Sample Documentation Form" teachers must provide documentation that the work sample demonstrates the state standards in writing. For each work sample, students must check one of the first three boxes on the form as well as the type of draft, process, and teacher-assisted for the work samples in the collection. The teacher must check that an "on-demand" essay is present in the collection. In the last box—teacher assistance—the student must describe what type of assistance he/she received beyond setting the prompt and the parameters of an effective paper.*
Reading Protocol	*In the "Work Sample Documentation Form" students and teachers must check all of the learning strands, both literary and informational. The student must provide of the titles of the texts must be provided to check the rigor of the readability of the texts. The student and the teacher must check each work sample to make sure that each sample addresses at least two strands. The student must identify which work sample is the short literary analysis paper and which is the short informational analysis paper. The teacher must check that an "on-demand" essay is present in the collection.*
Mathematics Protocol	*In the "Work Sample Documentation Form" students and teachers must check that all work samples address every high school content strand. Each work sample must address both a content strand and a process strand. Teachers must check that work samples meet the "rich problem" and high school level mathematics expectation. Students must check that each column and row have two entries. There must be an "on-demand" check*

Table 2D:

Scoring rules used to evaluate the collections – Performance criteria for the scoring rubrics used for each collection are given below along with an indication of the subject area EALRs and components within each EALR that are the focus of the performance criteria. Links to the EALRs are keys to authenticity validity.

Writing Criteria
Content, Organization & Style ➢ Has clear, focused main ideas or positions (EALR 1, Component 1) ➢ Elaborates by using reasons/arguments supported by well-chosen and specific details, examples, anecdotes, facts and/or statistics as evidence to support ideas or positions (EALR 1, Component 1) ➢ Includes information that is thoughtful and useful for the audience to know (EALR 1, Component 1) ➢ Organizes writing to make the best case to explain ideas or support positions (EALR 1, Component 2) ➢ Composes introductions that draw the reader into the main ideas or positions (EALR 1, Component 2) ➢ Writes conclusions that leave the reader with something to think about (EALR 1, Component 2) ➢ Organizes writing into effective, cohesive paragraphs (EALR 1, Component 2) ➢ Provides transitions which clearly serve to connect ideas (EALR 1, Component 2) ➢ Uses language effectively by exhibiting word choices that are effective and appropriate for intended audience, purpose, and form (EALR 1, Component 3) ➢ Writes (where appropriate) sentences or phrases that are varied in length and structure (EALR 1, Component 4) ➢ Provides the reader with a sense of the person behind the words (EALR 1, Component 5) Conventions ➢ Follows the rules of standard English [language] usage (EALR 1, Component 6) ➢ Spelling of commonly used words (EALR 1, Component 6) ➢ Capitalization (EALR 1, Component 6) ➢ Punctuation (EALR 1, Component 6) ➢ Exhibits the use of complete sentences except where purposeful phrases or clauses are used for effect (EALR 1, Component 6) ➢ Indicates paragraphs consistently (EALR 1, Component 6)
Reading Criteria
Comprehension of main ideas and details of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Identifies the main theme/main idea and uses evidence to demonstrate an overall understanding of the text (EALR 2, Component 1) ➢ Summarizes by providing an overarching statement about the text that connects to at least three events from the beginning, middle and end of text (EALR 2, Component 1) ➢ Infers and/or predicts about key elements of the text making connections with evidence (EALR 2, Component 1) ➢ Explains key vocabulary with both denotative and connotative definitions by linking them to the text (EALR 1, Component 2) Analysis, interpretation, & synthesis of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Applies knowledge of key literary/informational elements to enhance and expand understanding of text (EALR 2, Component 2) ➢ Compares and contrasts ideas to explain concepts within or between text (EALR 2, Component 3) ➢ Analyzes text to explain the relationship between cause(s) and effect(s) and links it back to the theme or main idea (EALR 2, Component 2) Thinks critically about literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Evaluate author's/ text's purpose and/or in order to judge effectiveness on intended audience ➢ Evaluates reasoning of ideas / themes within the text and makes connections with evidence Synthesizes information beyond the text by making generalizations, drawing conclusions, or applying information to evaluate a new text or context

Table 2D (Continued)

Mathematics Criteria
Uses high school content knowledge and procedures (EALR 1) with supporting work in: ➢ Number Sense (EALR 1, Component 1) ➢ Measurement (EALR 1, Component 2) ➢ Geometric Sense (EALR 1, Component 3) ➢ Probability & Statistics (EALR 1, Component 4) ➢ Algebraic Sense (EALR 1, Component 5) Solves Problems (EALR 2) ➢ Applies one or more strategies that lead to the answer (EALR 2, Component 2) ➢ Determines the answer to the problem (EALR 2, Component 3) Reasons Logically (EALR 3) ➢ Justifies conclusions, results, and/or answers by addressing the conditions and/or constraints in the problem Communicates Understanding (EALR 4) ➢ Gathers, represents, and/or shares mathematical information using clear mathematical language and organization Makes Connections (EALR 5) ➢ Uses and relates different mathematical models and representations of the same situation using clear mathematical language and organization (EALR 5, Components 1 and 2)

Table 2E

Range-Finding – The process of selecting exemplary collections to represent different performance levels

All Content Areas
Steps in the range-finding process ➢ Select a range of collections to serve as potential anchors for the rubrics during scoring training, practice collections to be used for practice during scoring training, and validity collections to be randomly inserted into scoring process to ensure adherence to scoring rubrics over time ➢ Ensure that all selected collections have met sufficiency criteria ➢ Discuss scoring rubrics ➢ Apply scoring rubrics to selected collections ➢ Discuss applied scores ➢ Adjust scoring rubrics and/or scores, if needed, based on collections ➢ Assign final scores to anchor collections ➢ Assign final scores to practice collections ➢ Assign final scores to validity collections

Table 2F

Scoring Training – The process of training scorers to apply scoring rubrics consistently using anchor collections to anchor rubrics

All Content Areas
Steps in the training process ➢ Review and discuss rubrics ➢ Review and discuss anchor collections ➢ Score practice collections ➢ Discuss assigned scores; work toward consensus with pre-assigned scores ➢ Score second practice collections ➢ Discuss assigned scores; work toward consensus with pre-assigned scores ➢ Scorers must qualify by meeting a criterion of exact agreement with pre-assigned scores

Table 2G

Table Scoring Process – The process of assigning scores to collections

All Content Areas
Steps in the scoring process ➢ Scorers assign scores ➢ Collections are randomly assigned to a second scorer (inter-rater agreement) ➢ Randomly selected collections are rescored by a table leader (supervisor) ➢ Validity collections are given to scorers randomly ➢ Scorers who drift from scoring rubrics are retrained as necessary

The next two tables, Tables 3A and 3B, link each of the Validity and Reliability standards COE design features.

Table 3A: Design Features of COE that Address Validity Standards

Validity Standard	Feature of COE Addressing Standard
Validity Standard 1: **Representation and Fidelity**	➢ Protocols for Reading, Writing, and Mathematics ➢ Sufficiency Review ➢ Scoring Rules ➢ Range-finding ➢ Scoring Training ➢ Scoring Process
Validity Standard 2: **Cognitive Demands**	➢ Protocols for Reading, Writing, and Mathematics
Validity Standard 3: **Consistency Across Assessments**	➢ Range-finding
Validity Standard 4: **Alignment with Instruction**	➢ Student self-report??
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	➢ Range-finding ➢ Scoring Training ➢ Scoring Process
Validity Standard 6: **Consequences of the Interpretation and Use of Assessment Results**	➢ Ongoing validity studies for the COE

Table 3B: Design Features of COE that Address Reliability Standards

Reliability Standard	Feature of COE Addressing Standard
Reliability Standard 1: **Generalizability**	➢ Protocols for Reading, Writing, and Mathematics
Reliability Standard 2: **Sufficiency of Evidence**	➢ Sufficiency Review ➢ Work Sample Documentation Form
Reliability Standard 3: **Clarity of Directions and Expectations**	➢ Protocols for Reading, Writing, and Mathematics ➢ Work Sample Documentation Directions ➢ Work Sample Sign-off Form
Reliability Standard 4: **Quality of Scoring**	➢ Scoring Rules ➢ Range-finding ➢ Scoring Training ➢ Scoring Process

References

Taylor, C. S. & Nolen, S. B. (1996). What does the psychometrician's classroom look like? *Educational Policy Analysis Archives*, v4, n17.

Taylor, C. S. & Nolen, S. B. (2005). *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms.* Columbus, OH: Pearson-Merrill-Prentice Hall.

Appendix A

Certificate of Academic Achievement (CAA) Options Advisory Committee Members

Linda Dobbs, Assistant Superintendent, ESD 189, Mt. Vernon, Washington

Deborah Gonzalez, Executive Director for Learning & Teaching, Puget Sound ESD, Highline, Washington

Gil Mendoza, Executive Director of Grants Management, Tacoma School District, Tacoma, Washington

Barbara Plake, Professor Emeritus, University of Nebraska-Lincoln

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

Edward Wiley, Professor, University of Colorado-Boulder

Appendix B

National Technical Advisory Committee for Assessment

Patricia Almond, University of Oregon, Eugene, Oregon

Peter Behuniac, University of Connecticut, Hartford, Connecticut

Richard Duran, Professor, California State University, Santa Barbara, California

George Englehard, Professor, Emory University, Atlanta, Georgia

Robert Linn, Professor Emeritus, University of Colorado-Boulder, UCLA-CRESST

William Mehrens, Professor Emeritus, Michigan State University, East Lansing, Michigan

Edys Quelmalz, Associate Director of the Center for Technology in Learning, Stanford Research Institute International, Palo Alto, California.

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington